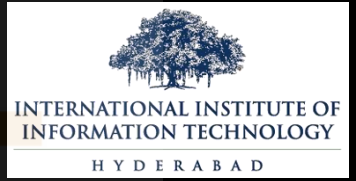
Responsible & Safe AI

Prof. Ponnurangam Kumaraguru (PK), IIITH
Prof. Balaraman Ravindran, IIT Madras
Prof. Arun Rajkumar, IIT Madras

RLHF, AI Alignment



RLHF: Intro

Reinforcement Learning Human Feedback

Optimizing for human preferences

Let's say we were training a language model on some task (e.g. summarization).

For an instruction x and a LM sample y, imagine we had a way to obtain a *human reward* of that summary: $R(x, y) \in \mathbb{R}$, higher is better.

x

```
SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco
...
overturn unstable
objects.
```

y_1

```
An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.
```

$R(x, y_1) = 8.0$

y_2

```
The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.
```

$R(x, y_2) = 1.2$

Now we want to maximize the expected reward of samples from our LM:

$$\mathbb{E}_{\hat{y} \sim p_\theta(y \mid x)}[R(x, \hat{y})]$$

High-level instantiation: 'RLHF' pipeline

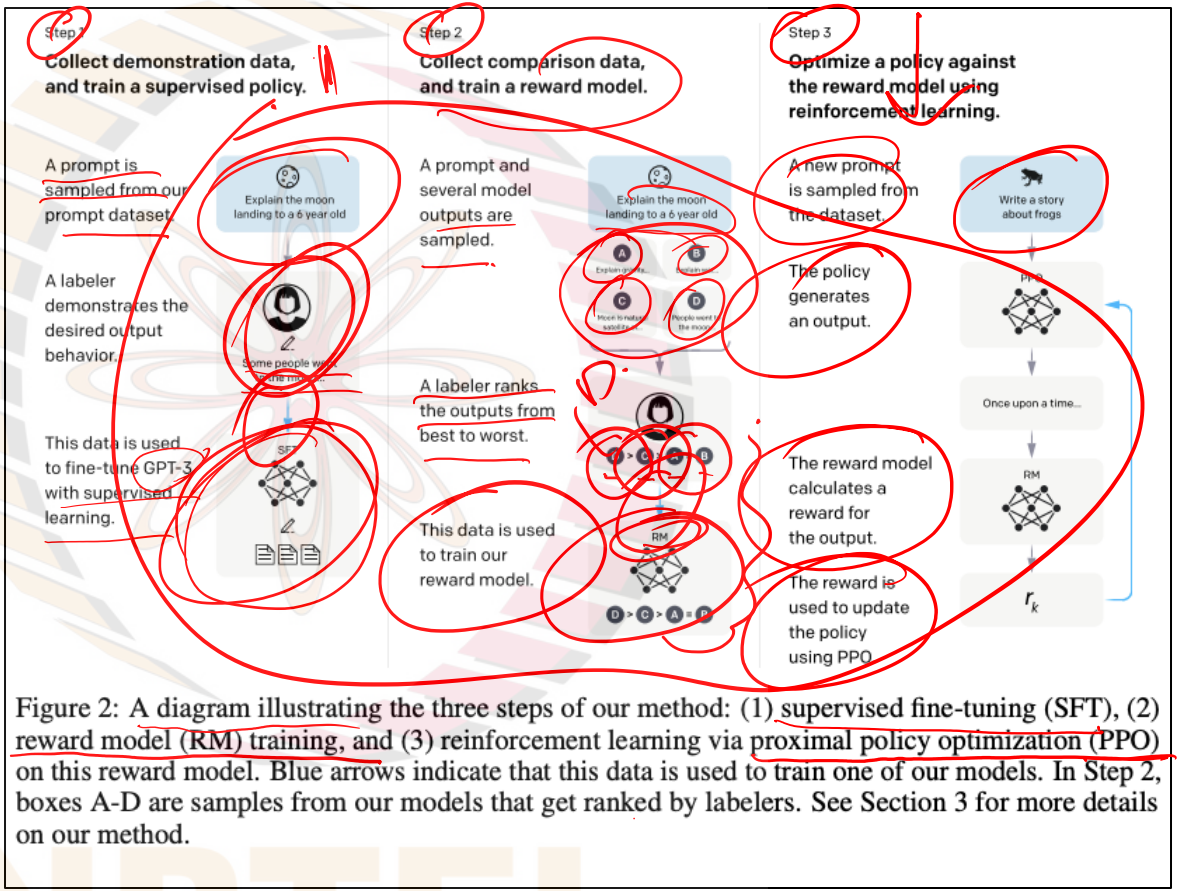


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

First step: instruction tuning!

Second + third steps: maximize reward

https://arxiv.org/pdf/2203.02155

How do we get the rewards?

Problem 1: human-in-the-loop is expensive!

Solution: instead of directly asking humans for preferences, model their preferences as a separate (NLP) problem!

An earthquake hit San Francisco. There was minor property damage, but no injuries.

The Bay Area has good weather but is prone to earthquakes and wildfires.

Train a $RM_\phi(x, y)$ to predict human reward from an annotated dataset, then optimize for RM_ϕ instead.

$R(x, y_1) = 8.0$ 

$R(x, y_2) = 1.2$ 

https://www.cs.utexas.edu/~pstone/Papers/bib2html-links/KCAP09-knox.pdf

How do we model human preferences?

Problem 2: human judgments are noisy and miscalibrated!

Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable

```
A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.
```

y_3

How do we model human preferences?

Problem 2: human judgments are noisy and miscalibrated!

Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.

y_3

$$R(x, y_3) = \quad 4.1? \quad 6.6? \quad 3.2?$$

https://escholarship.org/content/qt2nn523jz/qt2nn523jz.pdf
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190393

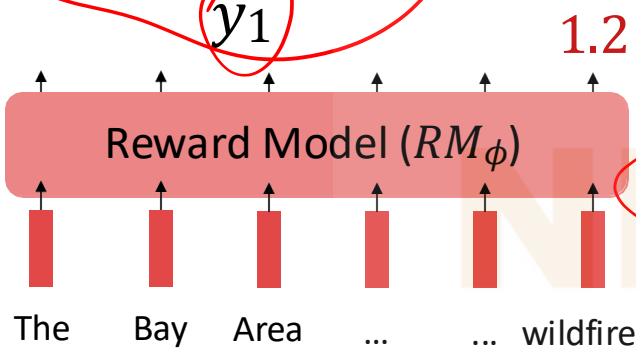
How do we model human preferences?

Problem 2: human judgments are noisy and miscalibrated!

Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable

An earthquake hit San Francisco. There was minor property damage, but no injuries. > A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage. > The Bay Area has good weather but is prone to earthquakes and wildfires.

y_1 y_3 y_2



Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(x, y^w, y^l)\sim D}\left[\log \sigma(RM_\phi(x, y^w) - RM_\phi(x, y^l))\right]$$

"winning" sample

"losing" sample

y^w should score higher than y^l

RLHF: Optimizing the learned reward model

We have the following:

- A pretrained (possibly instruction-finetuned) LM $p^{PT}(y \mid x)$
- A reward model $RM_\phi(x, y)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons

Now to do RLHF:

- Copy the model $p^{RL}_\theta(y \mid x)$, with parameters θ we would like to optimize
- We want to optimize:

RLHF: Optimizing the learned reward model

We want to optimize:

$$\mathbb{E}_{\hat{y}\sim p_\theta^{RL}(\hat{y}|x)}\left[RM_\phi(x,\hat{y})\right]$$

Do you see any problems?

Learned rewards are imperfect; this quantity can be imperfectly optimized

Add a penalty for drifting too for from the initialization:

$$\mathbb{E}_{\hat{y}\sim p_\theta^{RL}(\hat{y}|x)}\left[RM_\phi(x,\hat{y}) - \beta \log\left(\frac{p_\theta^{RL}(\hat{y}\mid x)}{p^{PT}(\hat{y}\mid x)}\right)\right]$$

Pay a price when $p_\theta^{RL}(\hat{y}\mid x) > p^{PT}(\hat{y}\mid x)$

This penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler (KL)** divergence between $p_\theta^{RL}(\hat{y}\mid x)$ and $p^{PT}(\hat{y}\mid x)$.

How to optimize? Reinforcement Learning!

RL has been around for a long time 1992

RL applied to deep-learning & game playing

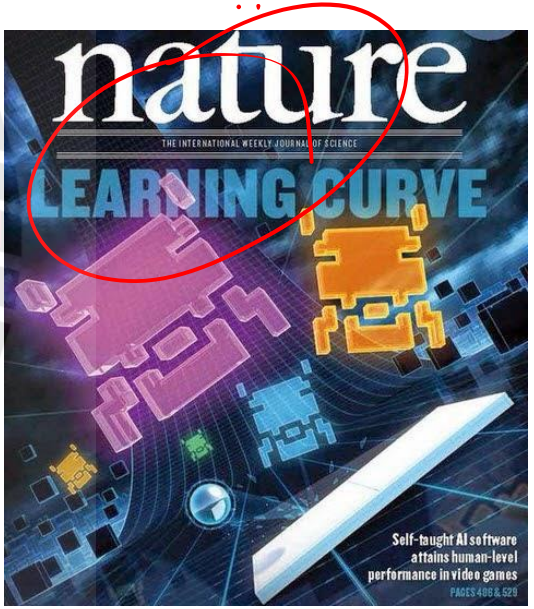
RL to LMs is new

RL in nutshell

Generate completions from p^{RL}_{θ} for several tasks

Compute reward using $RM_{\phi}(x, y)$

Update $p^{RL}_{\theta}(y \mid x)$ to increase probability of high-reward completions



AlphaGo

RLHF provides gains over pretraining + finetuning

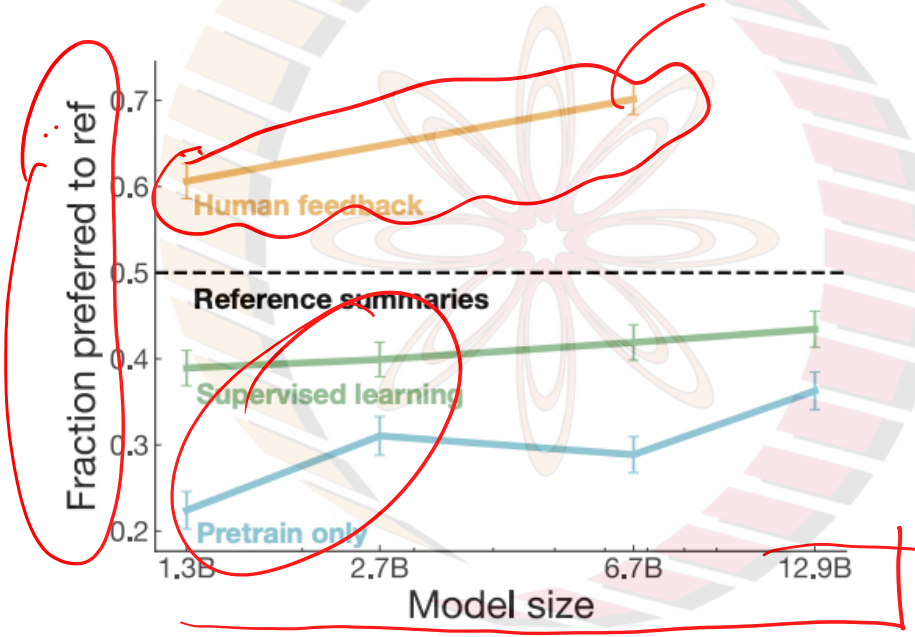


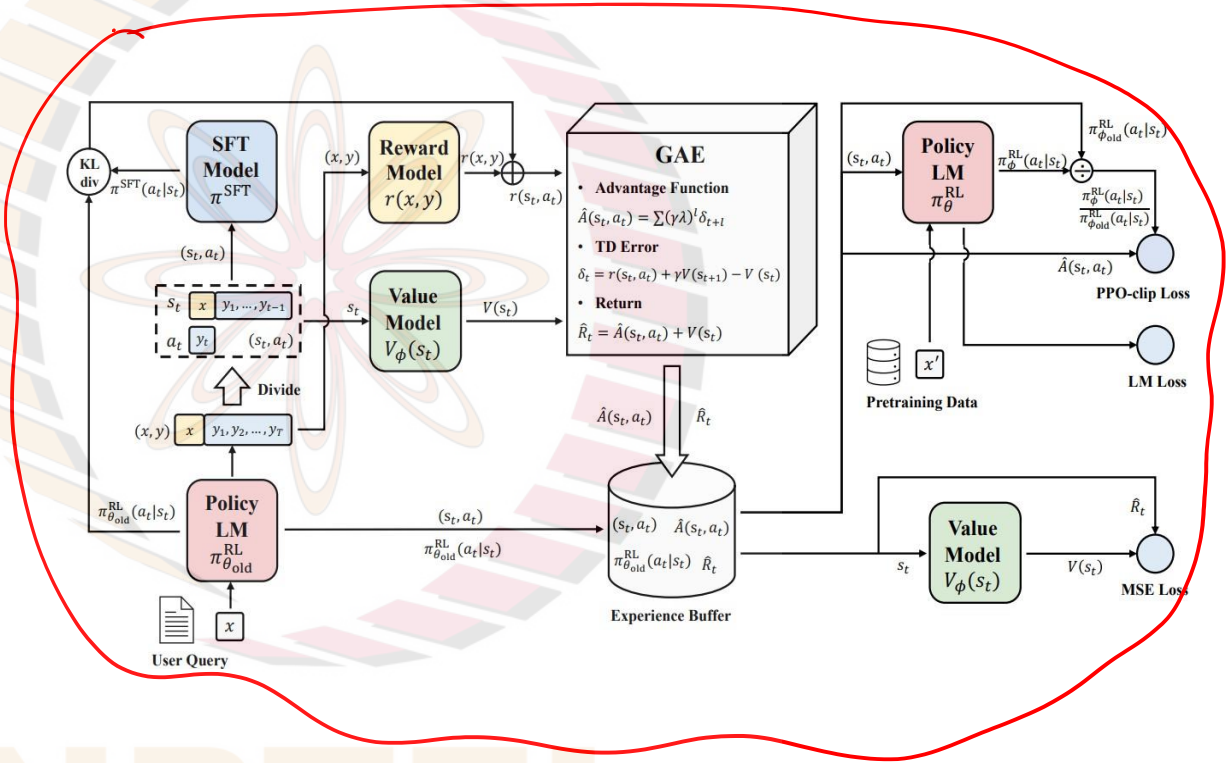
Figure 1: Fraction of the time humans prefer our models' summaries over the human-generated reference summaries on the TL;DR dataset.[4]Since quality judgments involve an arbitrary decision about how to trade off summary length vs. coverage within the 24-48 token limit, we also provide length-controlled graphs in Appendix F; length differences explain about a third of the gap between feedback and supervised learning at 6.7B.

https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

RLHF can be complex

RL optimization can be computationally expensive and tricky:

- Fitting a value function
- Online sampling is slow
- Performance can be sensitive to hyperparameters



https://arxiv.org/pdf/2307.04964

Summary RLHF

We want to optimize for human preferences

- Instead of humans writing the answers or giving uncalibrated scores, we get humans to rank different LM generated answers

Reinforcement learning from human feedback

- Train an explicit reward model on comparison data to predict a score for a given completion
- Optimize the LM to maximize the predicted score (under KL-constraint)
- Very effective when tuned well, computationally expensive and tricky to get right

InstructGPT: scaling up RLHF to large # of tasks

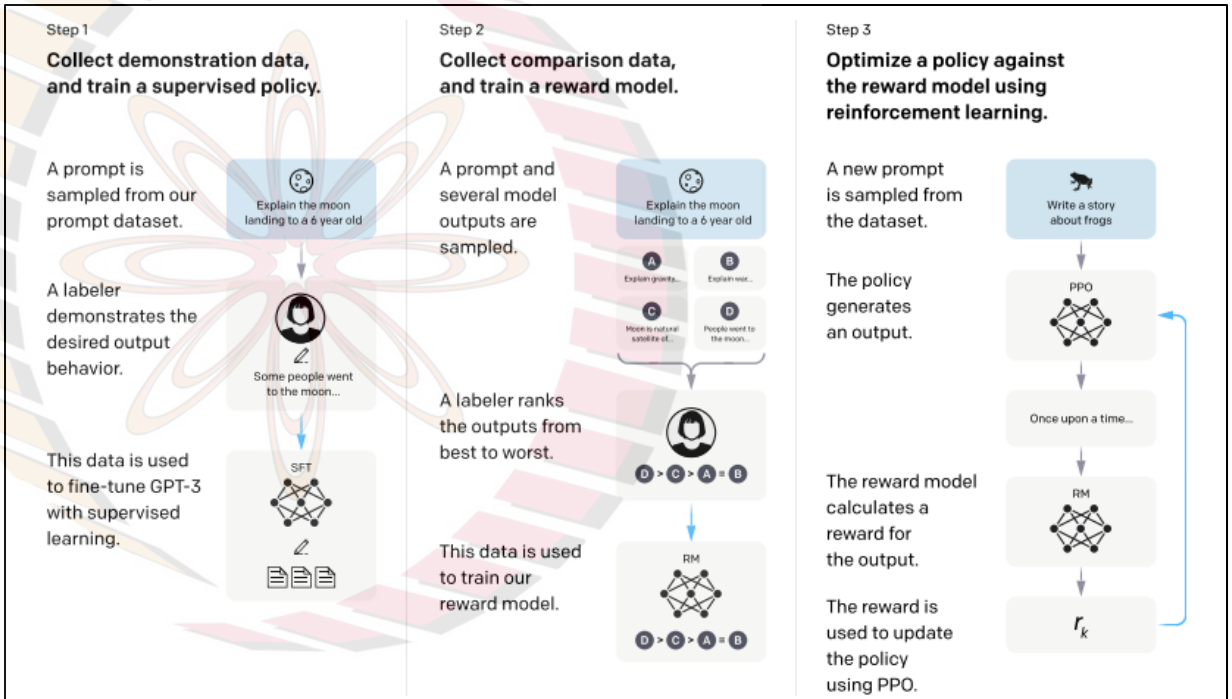


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

https://arxiv.org/pdf/2203.02155

InstructGPT: scaling up RLHF to large # of tasks

50K tasks

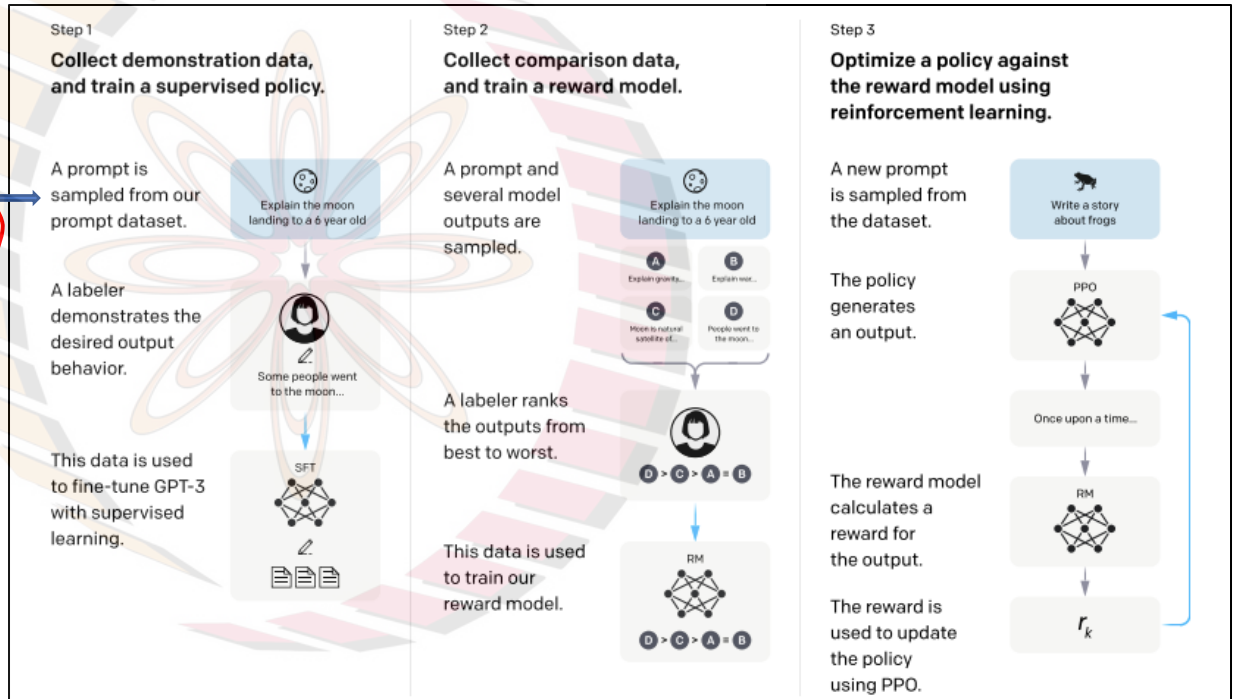


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

https://arxiv.org/pdf/2203.02155

InstructGPT: Kinds of prompts

Plain: arbitrary task, while ensuring the tasks had sufficient diversity

Few-shot: an instruction, and multiple query/response pairs for the instruction

User-based:

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.

https://arxiv.org/pdf/2203.02155

InstructGPT

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION

GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

InstructGPT

```
People went to the moon, and they took pictures of what they saw, and
sent them back to the earth so we could all see them.
```

InstructGPT

PROMPT *Write a short poem about a wise frog.*

COMPLETION

GPT-3

```
Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you
have lost.

Write a short story in which a character has a bad dream.
```

InstructGPT

```
The frog is a symbol of wisdom
He knows all the secrets of the world
He is a master of disguise
And a great teacher of life
He is a symbol of transformation
And the bringer of change
He is the frog who has seen it all
And knows the meaning of it all
```

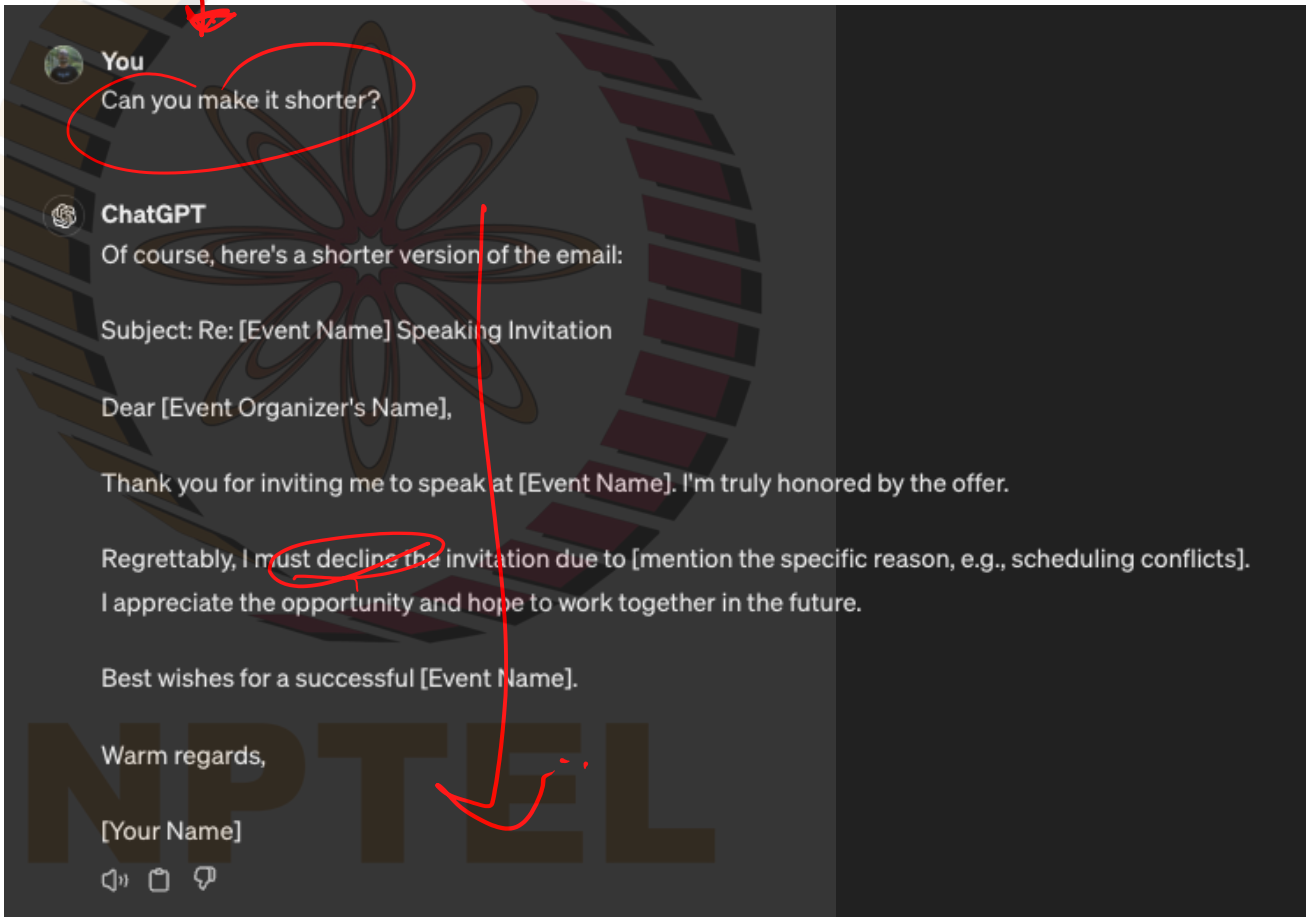
ChatGPT: Instruction Finetuning + RLHF for dialog agents

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

https://openai.com/blog/chatgpt/

ChatGPT


You
Can you make it shorter?
ChatGPT
Of course, here's a shorter version of the email:
Subject: Re: [Event Name] Speaking Invitation
Dear [Event Organizer's Name],
Thank you for inviting me to speak at [Event Name]. I'm truly honored by the offer.
Regrettably, I must decline the invitation due to [mention the specific reason, e.g., scheduling conflicts].
I appreciate the opportunity and hope to work together in the future.
Best wishes for a successful [Event Name].
Warm regards,
[Your Name]

Limitation of RL + Reward modeling

Rogue AI

Human preferences are unreliable
"Reward Hacking" is common problem in RL
To maximize the score
Win the game – Score (reward function)
did not matter
Agent found turbo boost, collect for more score
Large score, not doing race
Continuously crashing

We risk losing control over AIs as they become more capable
Proxy gaming: YouTube / Insta – User engagement – Mental health

Connection
Friendship
Insecurity
FOMO

What is an alignment problem?

What is an alignment problem?

Machine can solve problems that human will take decades to solve

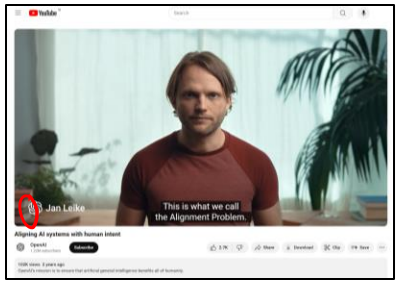
Sometimes less inclined towards listening to human instructions

Important to solve this problem to benefit AI effects

System should be incentivized to tell the truth (hallucination!)

Can the system act with human intentions & human values

Technically difficult to distinguish between AI generated & human generated (cool research problem!)



https://www.youtube.com/watch?v=yWDUzNiWPJA

RLHF for AI Alignment

Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback

Stephen Casper,* *MIT CSAIL,* scasper@mit.edu
Xander Davies,* *Harvard University*

Claudia Shi, *Columbia University*
Thomas Krendl Gilbert, *Cornell Tech*
Jérémy Scheurer, *Apollo Research*
Javier Rando, *ETH Zurich*
Rachel Freedman, *UC Berkeley*
Tomasz Korbak, *University of Sussex*
David Lindner, *ETH Zurich*
Pedro Freire, *Independent*
Tony Wang, *MIT CSAIL*
Samuel Marks, *Harvard University*
Charbel-Raphaël Segerie, *EffiSciences*
Micah Carroll, *UC Berkeley*

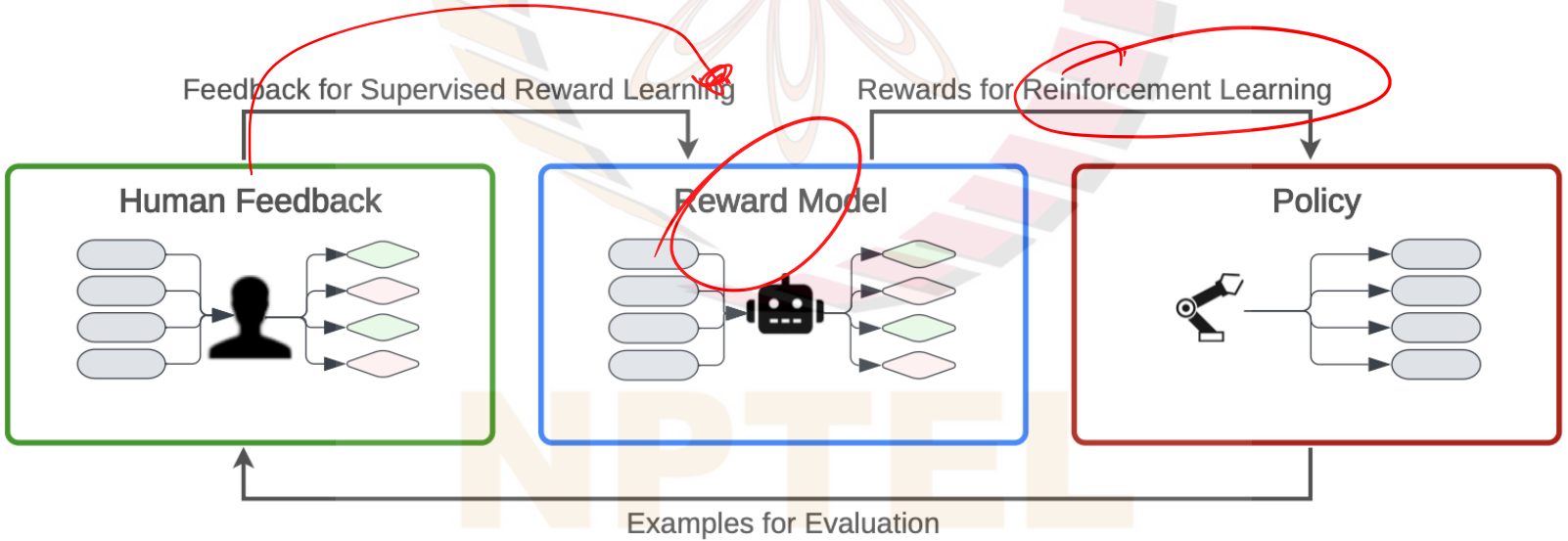
https://arxiv.org/abs/2307.15217

RLHF

RLHF – Human Feedback = ?

RLHF – Reward model = ?

RLHF – RL = ?

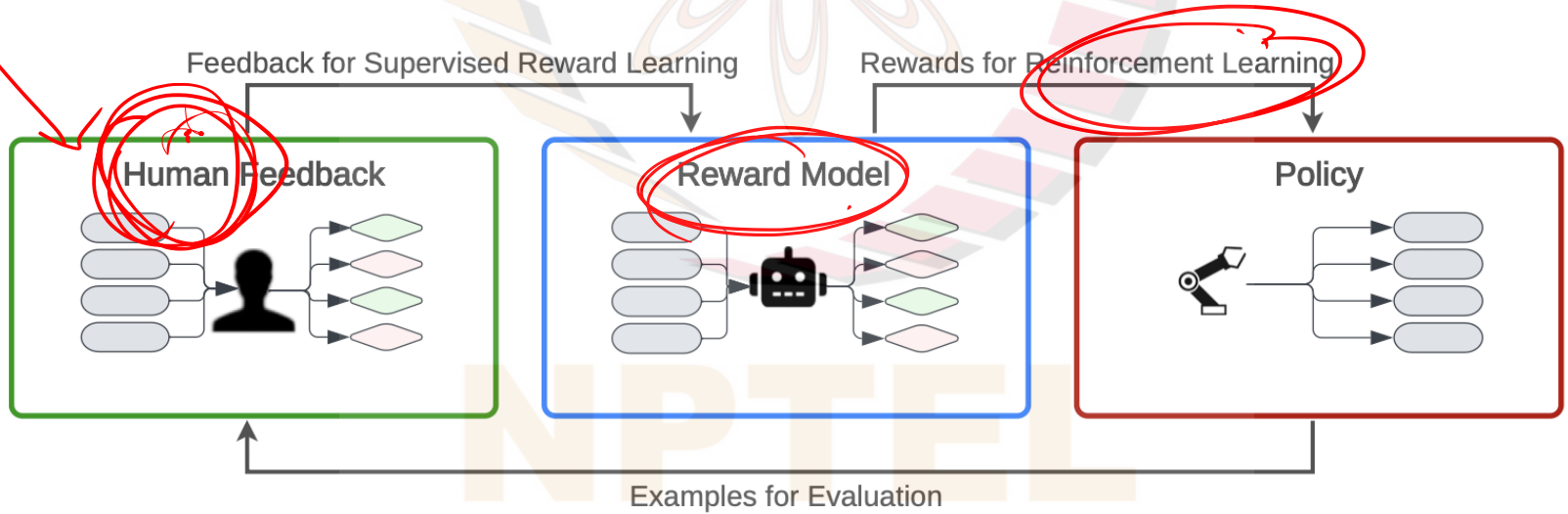


RLHF

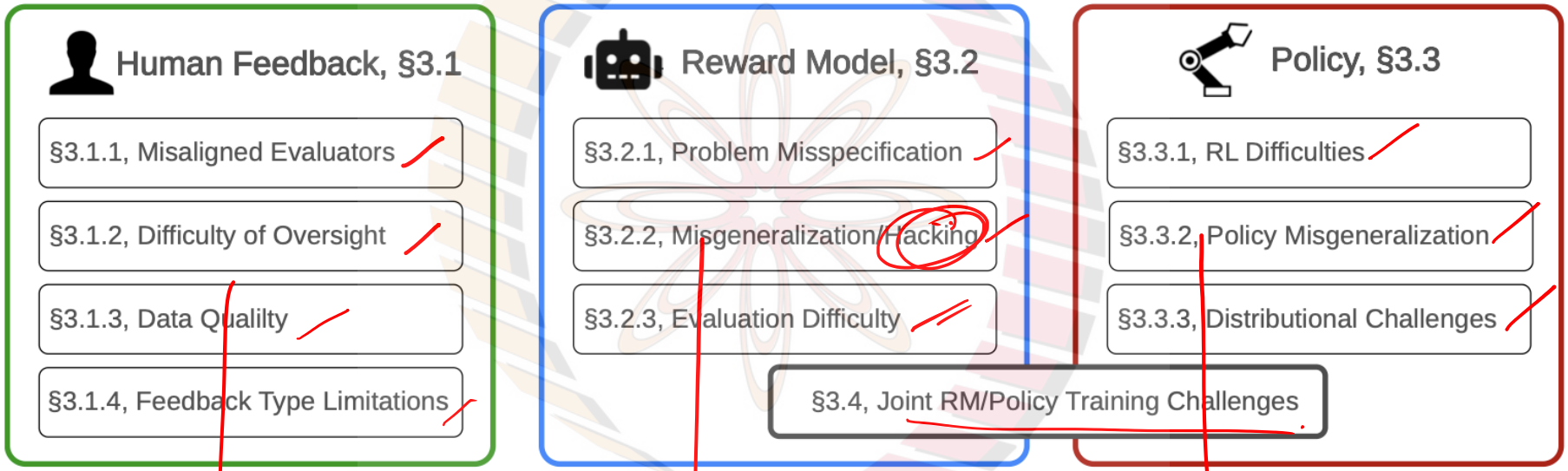
RLHF – Human Feedback = RLAIF (RL from AI Feedback)

RLHF – Reward model = Direct Preference Optimization (DPO)

RLHF – RL = Supervised Finetuning



RLHF Challenges



RLHF Challenges: Human Feedback

Selecting representative humans and getting them to provide feedback is difficult

Some evaluators have harmful biases and opinions

Individual human evaluators can poison data

Humans make simple mistakes due to limited time, attention, or care

Partial observability limits human evaluators

Humans can be misled, so their evaluations can be gamed

Data collection can introduce harmful biases

Suffers from trade-off between richness and efficiency of feedback types (comparison, scalar, label, etc.)

RLHF Challenges: Reward model

An individual human's values are difficult to represent with a reward function

A single reward function cannot represent a diverse society of humans

Optimizing for an imperfect reward proxy leads to reward hacking

Evaluating reward model is difficult and expensive

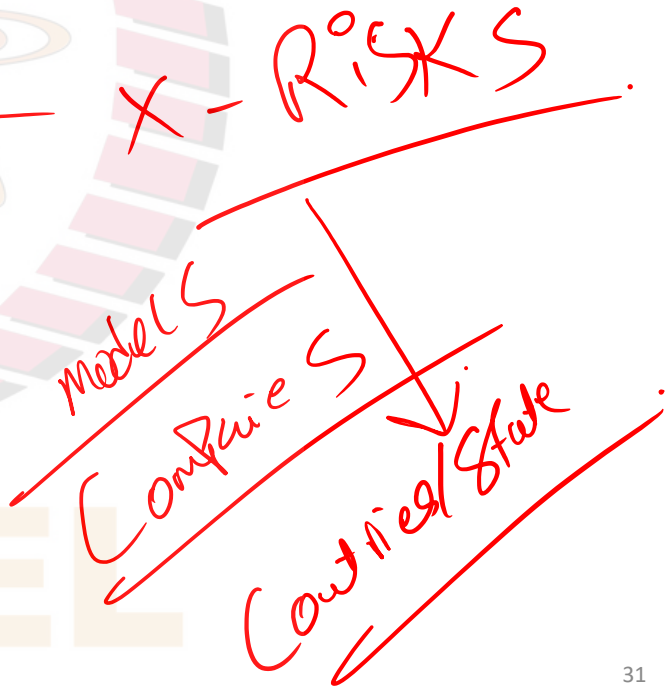
RLHF Challenges: Policy

Policies tend to perform poorly in adversarial situations

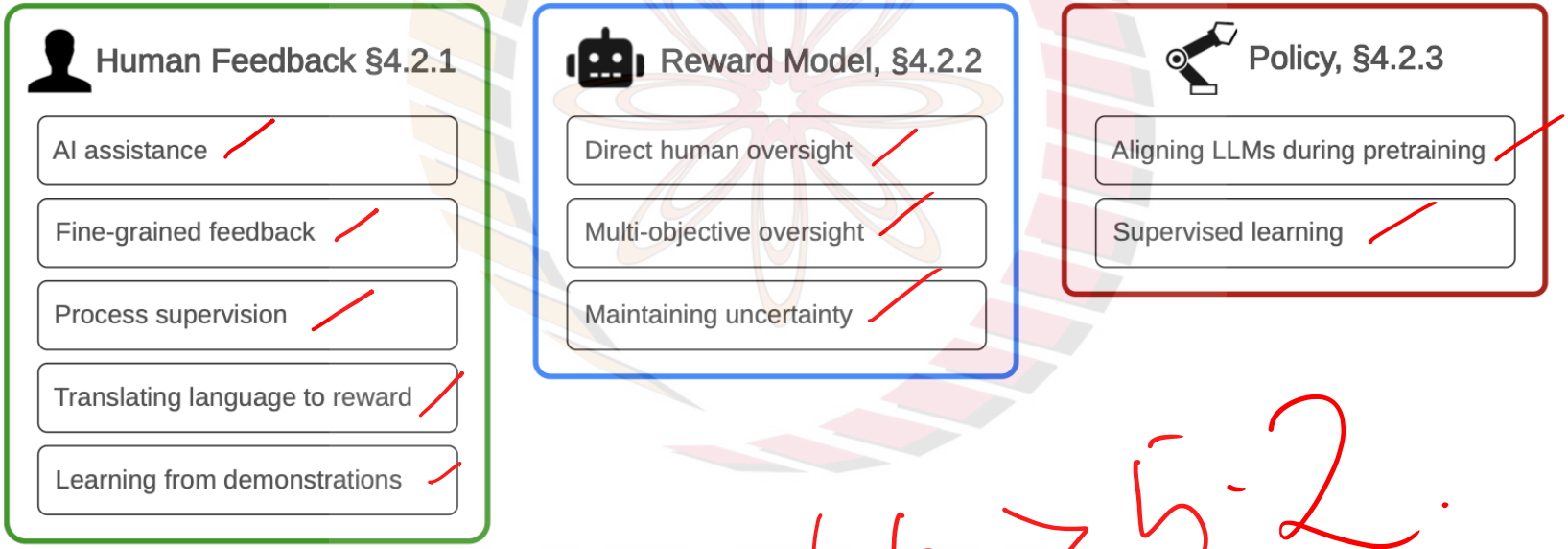
Policies tend to perform poorly in deployment even if rewards seen during training were perfectly correct

Optimal RL agents tend to seek power

RL contributes to mode collapse

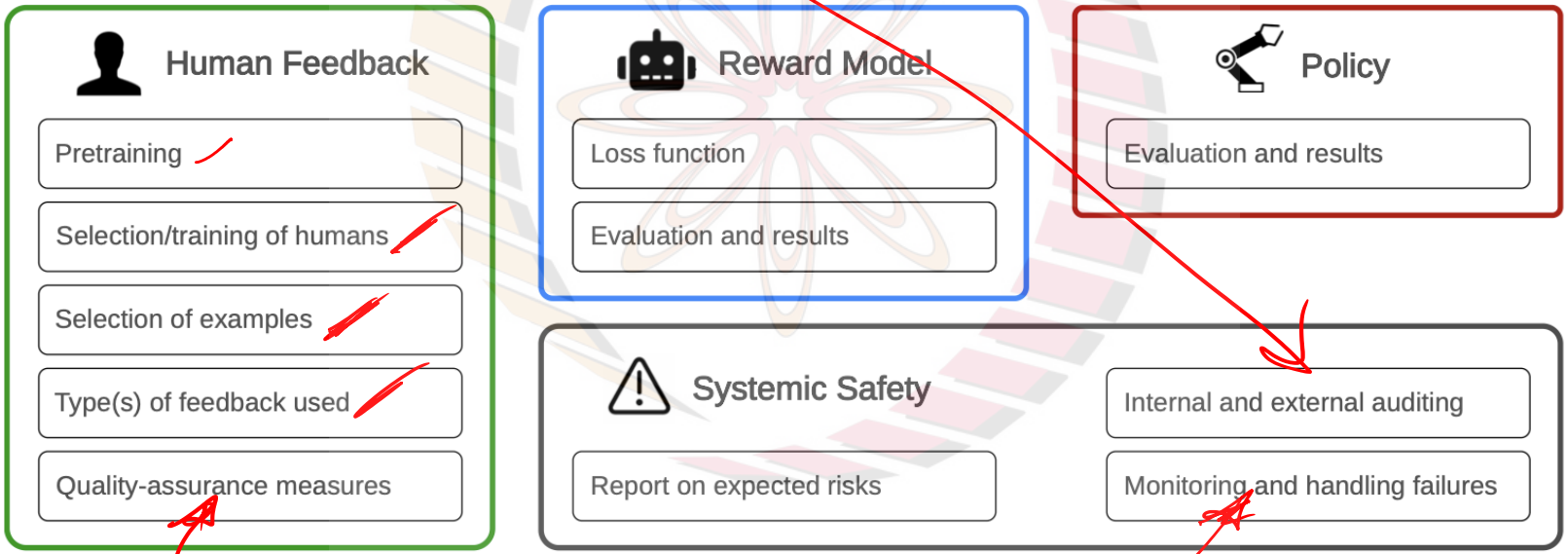


Strategies to address the RLHF challenges



4.6 > 5.2.

Transparency / Auditing for RLHF



Auditing

Black-Box Access is Insufficient for Rigorous AI Audits

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, Dylan Hadfield-Menell

External audits of AI systems are increasingly recognized as a key mechanism for AI governance. The effectiveness of an audit, however, depends on the degree of access granted to auditors. Recent audits of state-of-the-art AI systems have primarily relied on black-box access, in which auditors can only query the system and observe its outputs. However, white-box access to the system's inner workings (e.g., weights, activations, gradients) allows an auditor to perform stronger attacks, more thoroughly interpret models, and conduct fine-tuning. Meanwhile, outside-the-box access to training and deployment information (e.g., methodology, code, documentation, data, deployment details, findings from internal evaluations) allows auditors to scrutinize the development process and design more targeted evaluations. In this paper, we examine the limitations of black-box audits and the advantages of white- and outside-the-box audits. We also discuss technical, physical, and legal safeguards for performing these audits with minimal security risks. Given that different forms of access can lead to very different levels of evaluation, we conclude that (1) transparency regarding the access and methods used by auditors is necessary to properly interpret audit results, and (2) white- and outside-the-box access allow for substantially more scrutiny than black-box access alone.

https://arxiv.org/abs/2401.14446

State of Alignment

RLHF was never designed to solve the AI alignment problem

It is probably the state of art for now for alignment problem

RLHF Bibliography

https://web.stanford.edu/class/cs224n/slides/cs224n-spr2024-lecture10-prompting-rlhf.pdf


pk.profgiri


Ponnurangam.kumaraguru


/in/ponguru


ponguru


pk.guru@iiit.ac.in

Thank you for attending the class!!!